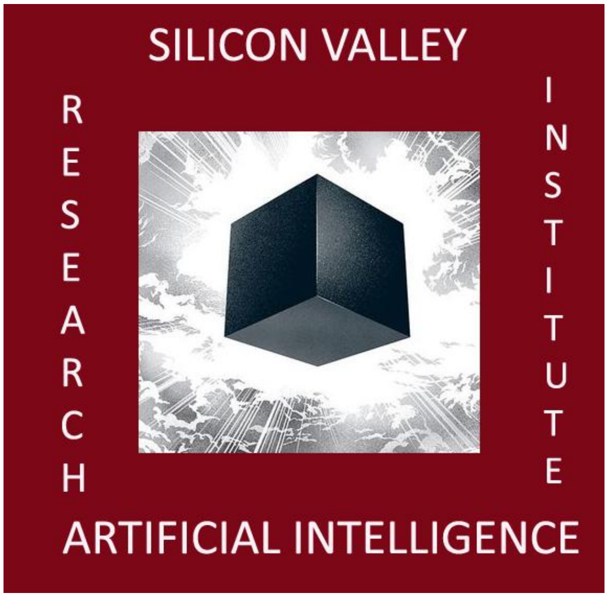

A.I. for Theory Formation

It is an old ambition of A.I. to simulate intuition and ingenuity in a computer program.

After Douglas Lenat's half-baked attempt at Stanford to create a program for mathematical discovery, Automated Mathematician or AM (1976), Herbert Simon's students at Carnegie Mellon debuted discovery systems that employed knowledge ("heuristics") about a specific domain to derive scientific laws from data: Fahrenheit in chemistry, a system designed by Jan Zytkow ("A Theory of Historical Discovery", 1986); Kekada in biology, by Deepak Kulkarni ("The Processes of Scientific Discovery", 1988); and Mechem also in chemistry, by Raul Valdes-Perez (1990, that includes the hypothesis-formation program Stoich). More general was Abacus, an evolution of Ryszard Michalski's AQ11, that was applied to problems both in physics and chemistry ("Integrating Quantitative and Qualitative Discovery", 1986). Kulikowski's student Vonwun Soo at the University of Pennsylvania worked on chemical reaction pathways before Kekada and Mechem ("Theory Formation in Postulating Enzyme Kinetic Mechanisms", 1987). In 1988 Siemion Fajtlowicz at the University of Houston developed Graffiti for discovery in mathematics. Peter Karp at Stanford University wrote the program Hypgene for hypotheses generation in molecular biology (1989), an evolution of Molgen.

Douglas Hofstadter at Indiana University built several programs such as Jumbo (1983) and Numbo (1987) that reacted against the notion that intelligence could be just the product of domain knowledge. He believed that finding patterns constituted the core of intelligence and therefore researched a more abstract type of intelligence, capable of discovering patterns and shaping concepts out of patterns. He implicitly viewed mathematics as the supreme demonstration of human intelligence because mathematicians continuously discover higher and higher levels of abstraction to explain the patterns that they discover. Hofstadter's programs tried to simulate how we build concepts, expand them, adapt them. Concepts are dynamic, not static, or, better, concepts are fluid. An important moment is when a paradigm shift occurs and we start seeing the world differently. His student Melanie Mitchell built Copycat (1988) to simulate paradigm shifts in a microworld. Note that most research in mathematics is driven by the illusion of getting closer and closer to the ultimate truth of the universe, not by the certainty of doing something useful for daily life.

Ken Forbus's student Brian Falkenhainer at Northwestern University used Forbus' analogical reasoning system SME to develop scientific theories ("Scientific Theory Formation through Analogical Inference", 1987).

Saul Amarel's student Michael Sims at Rutgers University built the program IL for theory formation in mathematics (1990). In 1997 Derek Sleeman's student Faye Mitchell at Aberdeen University in Britain developed Daviccand (Data VIsualisation Clustering and Conceptually Analysing) for chemical properties of metals (and the rare case of interactive human-machine discovery). Alan Bundy's student Simon Colton at the University of Edinburgh designed the HR system for theory formation in mathematics (1998).

Meanwhile, the field of "literature-based knowledge discovery" had been launched by Don Swanson at the University of Chicago in 1986 ("Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge", 1986). Swanson viewed discovery as connecting the disconnected in scientific knowledge, as finding links between apparently unrelated ideas in the scientific literature. Swanson, who knew little or nothing about fish oil and a vascular disease called Reynaud's Syndrome, discovered that fish oil can cure that disease without conducting any experiment. He later also discovered a link between migraine and magnesium deficiency. Swanson's idea was relatively simple: if a discipline is studying the relationship between concept A and B and another discipline is studying the relationship between concepts B and C, there might be important hidden facts about the relationship between A and C that are not visible to either discipline. The catch, of course, is that different disciplines use different vocabularies, so one has to translate one field's terminology into the other field's terminology. Neil Smalheiser then helped him develop Arrowsmith, a program that scanned the scientific literature to find connections between articles (1998).

In 1997 Marti Hearst started the LINDI (Linking Information for Novel Discovery and Insight) project at UC Berkeley (later renamed Bio Text).

These programs for scientific discovery did not complete the loop from hypothesis generation to experiment design back to hypothesis generation. In 2004 Ross King at Aberystwyth University envisioned a "robot scientist", Adam, designed to: generate hypotheses, design experiments to test these hypotheses, carry out the physical experiments using robotic laboratories, interpret the resulting data, and repeat the cycle until a successful theory is shaped ("Functional Genomic Hypothesis Generation and Experimentation by a Robot Scientist", 2004). In 2009 Adam identified twelve genes responsible for catalysing specific chemical reactions in yeast.

Eureqa, developed in 2007 by Hod Lipson's student Michael Schmidt at Cornell University (later a product sold by Schmidt's startup Nutonian), uses evolutionary algorithms to generate and select hypotheses consistent with experimental data and to design further experiments to continue the selection.

In 2009 Andrey Rzhetsky's team at the University of Chicago, drawing data from more than 300,000 papers and 8 million abstracts, built two datasets of molecular interactions related to the muscle disease ataxia: 49,493 for mice and 52,518 for humans. By comparing the two, the scientists were able to discover genes associated with brain malformations ("Looking at Cerebellar Malformations through Text-Mined Interactomes of Mice and Humans", 2009). Rzhetsky later wrote the manifesto "Machine Science" (2010).

Pierre-Yves Oudeyer at Inria (the French Institute for Research in Computer Science and Automation) worked on computational models of curiosity ("Intrinsic Motivation Systems for Autonomous Mental Development", 2007).

Joshua Tenenbaum, Charles Kemp (now at CMU) and Thomas Griffiths (now at UCB) proposed that hierarchical Bayesian models are the force behind our brain's ability to discover scientific theories ("A Probabilistic Model of Theory Formation", 2009).

IBM Research in San Jose and Baylor College of Medicine in Texas collaborated on KnIT, the Knowledge Integration Toolkit (2014), a system that mines scientific literature and generates hypotheses ("Automated Hypothesis Generation Based on Mining Scientific Literature"). KnIT read 100,000 papers and discovered nine p53 kinases, seven of which had already been discovered by scientists and two that were unknown.

In 2015 a program developed by Michael Levin's group at Tufts University developed a scientific theory about a 120-year-old mystery related to the flatworm planarian; a modest but encouraging success.

These were all systems that used preexisting knowledge to analyze data and generate new knowledge. Since 2006, however, the ruling paradigm in A.I. was "deep learning", i.e. neural networks, i.e. data-driven A.I. Neural networks discover patterns, not theories. Can one discover new knowledge in a field without any previous knowledge of the field?

In June 2017, at the 50th Alan Turing Award ceremony in San Jose, Stuart Russell of UC Berkeley declared "A deep-learning system would never discover the Higgs boson from the raw data". Unbeknownst to him, two months earlier DeepMind had just seen a program discover something that was not in the data, AlphaGo Zero. So the verdict is still out on whether a deep-learning system can or cannot discover something like the Higgs boson.

Relatively few scientists followed Solomonoff's third way to artificial intelligence. Marcus Hutter of the Australian National University worked out a version of Solomonoff induction called AIXI, a universal optimal learning program, which he claimed to be "the most intelligent unbiased agent possible" ("A Theory of Universal Artificial Intelligence based on Algorithmic Complexity", 2000). He then (2009) used a Monte Carlo method (that randomly generates the set of hypotheses) to build a a computationally feasible approximation.

In 2002 Juergen Schmidhuber at IDSIA designed an Optimal Ordered Problem Solver (OOPS) and then in 2003 a self-improving problem solver called Goedel Machine. These programs modify themselves when they prove logically that a change improves their performance. (Schmidhuber believes that the history of Artificial Intelligence begins with Kurt Goedel's famous self-referential formulas of 1931).

One foundational problem is that we don't quite understand well how human creativity works. One of the most famous theories was advanced by Robert Sternberg at Yale University ("A Propulsion Model of Types of Creative Contributions", 1999). His definition was that creative contributions propel a field forward, and he recognized eight types of such contributions because theory formation is not a well-defined process, nor is it a single process. Sternberg's paper "Types of Innovation" (2003) is a good summary of famous discoveries classified according to their different dynamics: replication, redefinition, forward incrementation, redirection, reinitiation, integration, etc. Some creative contributions accept current paradigms, some reject them and some integrate multiple paradigms. His collaborators James Kaufman and Lauren Skidmore provided an updated survey for the Internet age in "Taking the Propulsion Model of Creative Contributions into the 21st Century" (2010).